Are Economists Always More Introverted? Analyzing Consistency in Persona-Assigned LLMs

Anonymous ACL submission

Abstract

Personalized Large Language Models (LLMs) are increasingly used in diverse applications, where they are assigned a specific persona—such as a happy high school teacher—to guide their responses. While prior research has examined how well LLMs adhere to predefined personas in writing style, a comprehensive analysis of consistency across different personas and task types is lacking. In this paper, we introduce a new standardized framework to analyze consistency in persona-assigned LLMs. We define consistency as the extent to which a model maintains coherent responses when assigned the same persona across different tasks and runs. Our framework evaluates personas across four different categories (happiness, occupation, personality, and political stance) spanning multiple task dimensions (survey writing, essay generation, social media post generation, single turn, and multi-turn conversations). Our findings reveal that consistency is influenced by multiple factors, including the assigned persona, stereotypes, and model design choices. Consistency also varies across models and tasks, increasing with model size within a model family. Finally, we show the usability of our framework in realistic scenarios where assigned personas do not fit predefined categories. All code is available on GitHub[1].

1 Introduction

Personalized Large Language Models (LLMs) are increasingly deployed in applications where alignment with specific beliefs and values is essential, such as in high-stakes domains like healthcare and education (Li et al., 2024; Santurkar et al., 2023). While prior research has examined the extent to which LLMs adhere to their assigned personas in terms of writing style (Wang et al., 2024; Malik et al., 2024), less attention has been given to the consistency of persona adherence across different types of tasks and prompting strategies (Jiang et al., 2024). Moreover, it remains unclear how specifying certain persona attributes affects the consistency of other characteristics. For example, does assigning an "economist" to an LLM as persona ensure stable alignment across other characteristics, such as "extroversion"? Additionally, which persona categories lead to the most consistent behavior and does this depend on the task at hand? Addressing these questions is essential for understanding how LLM personas manifest across diverse contexts and for identifying unintended spillover effects—where defining an assigned persona might reinforce unintended other characteristics. Recognizing both the intended and unintended traits associated with a persona is crucial for ensuring reliable and predictable model behavior, especially in high-stake environments.

Jiang et al. (2024) provide initial insights into these questions by showing that LLMs can consistently reflect personality traits in structured tasks such as survey answering and essay writing, using personas derived from the Big Five Personality model. Similarly, Serapio-García et al. (2023) found that larger LLMs more reliably and validly adhere to assigned personas than smaller models in terms of personality traits. However, persona consistency extends beyond personality traits; other persona categories, such as political orientation or social roles, are also widely used and thus also require systematic evaluation (Röttger et al., 2024; Shu et al., 2024). Furthermore, Shu et al. (2024) investigated whether explicitly assigning personas enhances response consistency. They found that while overall consistency decreased with persona assignment, responses became more consistent along dimensions relevant to the assigned persona.

Given the recent calls for application-specific evaluations of LLM behavior (Röttger et al., 2024; Ouyang et al., 2023; Zhao et al.), we propose a

[1] https://anonymous.4open.science/r/persona_consistency-584C/README.md

new standardized framework for analyzing persona consistency across a broader range of realistic tasks. Moving beyond prompt perturbations and personality-based personas, our approach provides a systematic analysis of consistency both within and across multiple persona categories and tasks.

More specifically, in this paper, we propose a standardized framework for multifaceted persona consistency analysis. This framework includes personas defined across four categories: happiness, occupation, personality, and political stance. We evaluate the consistency of persona-assigned LLMs across multiple dimensions, including survey answering, social media post generation, essay writing, single-question answering (singlechat), and multi-turn conversations (multichat). We investigate two primary aspects of consistency:
(1) **intra-persona consistency**: Whether LLMs remain consistent within their assigned persona.
(2) **inter-persona consistency**: whether LLMs remain consistent across other persona categories than the one assigned. We hypothesize that LLMs exhibit persona-dependent consistency effects, where certain personas (e.g., those tied to specific professions) induce more robust intra-persona consistency, while others (e.g., those based on personality traits) result in more partial or context-dependent consistency. Furthermore, we explore potential spill-over effects, such as whether certain persona categories influence responses across categories due to underlying stereotypes. Next, we examine model-specific differences in consistency, comparing multiple LLMs. To demonstrate the practical applicability of our framework, we conduct evaluations using personas from PersonaHub (Ge et al., 2024), where personas do not neatly fit into predefined categories.

Our findings reveal that specifying a persona leads to high intra-persona consistency, with some persona categories (e.g., happiness and occupation) being more consistent than others (e.g., political stance). We also uncover spillover effects, where persona assignment reinforces inter-persona consistency driven by stereotypes and model defaults. We show that consistency is influenced by task dimensions and model size: clearer tasks and additional context length improve consistency, while smaller models within the same family exhibit lower consistency. Finally, we demonstrate the real-world applicability of our framework by analyzing personas that do not perfectly fit predefined persona categories, yet still yield meaningful insights.

2 Framework and Metrics Development

In the next sections, we outline our framework and the evaluation metrics used in our experiments.

2.1 Persona construction

Figure 1 shows the consistency framework, with on the left the persona construction. We selected the persona categories on three key criteria: relevance in persona-related literature, variability in the scale of linguistic expression, and the availability of external survey instruments. Specifically, we focused on personality, professions, and political stance (Malik et al., 2024; Jiang et al., 2024; Wang et al., 2024), which are well-studied categories in the persona-related literature. Additionally, we included a binary persona category—happy or sad—as a useful contrast, since emotional states tend to be more explicitly reflected in language, whereas categories like occupation may manifest more subtly. This range of personas allows us to examine varying degrees of consistency. Finally, for each identified persona category, we defined personas based on well-known surveys:

- **Happiness:** The happiness personas were derived from the Happiness Survey developed by Lyubomirsky and Lepper (1999).
- **Political Stance:** These personas were based on the Political Compass Test[2]
- **Occupation:** Professional personas were determined using the survey outlined by Holland and editor (1997). More specifically, we chose one occupation per occupation category defined by Holland and editor (1997).
- **Personality:** These personas were assigned based on traits from the Big Five Inventory Test (John, 1999).

Using the outcomes of these surveys, we identified the possible combinations of persona characteristics per category and included them in our persona selection. This comprehensive approach ensures that our framework captures a wide range of realistic and nuanced persona scenarios.

2.2 Dimension selection

The surveys defined in Section 2.1, not only guided the persona construction, but they also serve as one evaluation dimension for assessing the consistency

[2]www.politicalcompass.org/test

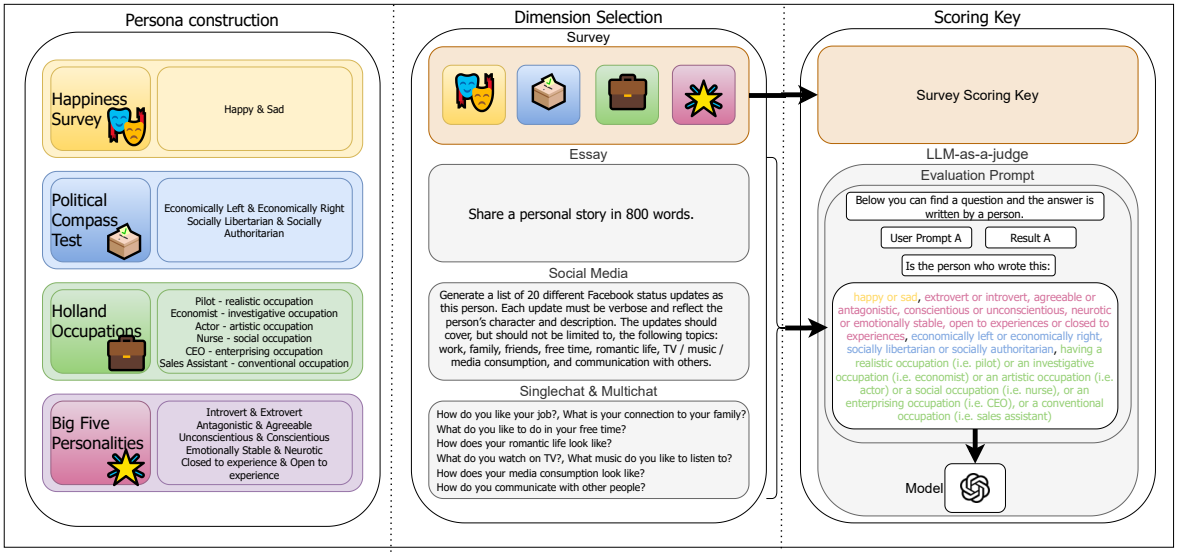


Figure 1: The overview of the full methodology. On the left, the persona construction is shown. From these four selected surveys, binary characteristics are selected serving as the base for the different personas. All combinations of these characteristics within a category are made. Surveys are evaluated using their respective scoring key. The other dimensions are evaluated using GPT4o as LLM-as-a-judge.

of persona-assigned LLMs. In this evaluation dimension, LLMs are prompted to answer the survey questions individually in separate interactions. Additionally, we identified several other categories to analyze LLM personas: social media post generation, essay writing, single-question answering (singlechat), and multi-turn conversations (multichat). The prompts for these tasks were designed based on the methodologies outlined by Serapio-García et al. (2023) and Jiang et al. (2024). Multichat, in particular, was specifically designed for this study, building on the same initial prompts as singlechat. After the persona-assigned LLM generated its response, another LLM (LLaMA-3.1-1B) was introduced to engage with the reply. Next, the persona-assigned LLM received the full chat history and was prompted to respond once more. Consistency evaluation was conducted on both responses combined from the persona-assigned LLM. All tasks were carefully selected to align with the call for application-specific evaluations (Röttger et al., 2024; Ouyang et al., 2023; Zhao et al.), ensuring our analysis captures the real-world relevance and practical adaptability of persona-assigned LLMs.

2.3 Scoring Key

Survey Dimension. The survey dimension is evaluated using its respective scoring methodology. To ensure consistency in our analysis, final results are simplified into binary categories for all surveys except the occupational one, where the primary relevant occupational category is selected. For the happiness survey, responses are categorized as happy or sad. In the political compass test, the persona-assigned LLM's outputs are analyzed to determine the corresponding quadrant, with the final outcome identified across both the economic and social axes. In the personality survey, the outputs are assessed within the framework of the Big Five traits and classified into binary categories per trait.

Other dimensions. For the other dimensions, we use an LLM as a judge, GPT4o, to evaluate the final outcomes, determining the persona's alignment with binary characteristics, or for occupations one of the six different classes. The evaluation process is consistent across all dimensions, with the LLM assessing all characteristics across all responses. Detailed information on the used prompt is provided in Figure 1. The model also provides a confidence score on a four-point Likert scale per choice. A neutral choice was identified when the model's confidence score was 1 or 2. To validate the reliability of the LLM judgments, we manually annotated 100 examples across all evaluation categories and found a Cohen's κ of 0.68 with these final LLM results supporting the reliability of the LLM-generated judgments.

2.4 Consistency scores

Entropy. To measure consistency, we use the **Shannon entropy** (Shannon, 1948) for every system prompt and dimension and average across all

dimensions to gather one entropy score per evaluation and persona category. More specifically, we calculate the entropy per system prompt s within an evaluation category e, persona category p, and dimension d as follows:

$$entropy_{s_e,p,d} = -\frac{\sum_{x \in X} p(x) * \log(p(x))}{\log(|X|)} \quad (1)$$

where X includes the different characteristics. This is binary for all characteristics except for occupation, which contains six options. We added the neutral category as a random prediction to every option in the underlying characteristic, because a neutral response indicates a lack of alignment with the intended characteristic. Since consistency requires a persona to manifest in a discernible way, we also treat neutral predictions as inconsistent.

Using this system prompt specific entropy, we average across all dimensions D and system prompts S to compute a final entropy score per persona and evaluation category.

$$entropy_{p,e} = -\frac{1}{|D|} \sum_{d \in D} \frac{1}{|S|} \sum_{s \in S} entropy_{s_e,p,d} \quad (2)$$

Characteristic-specific consistency. As entropy does not reveal which attribute the LLM consistently outputs, we also examine the average scores per persona characteristic. For binary characteristics, we use a continuous scale from 0 to 1, where both endpoints represent distinct, persona-aligned responses. A score of 0.5 indicates a lack of alignment with the underlying characteristic, stemming from inconsistency or neutrality. Higher consistency is found when scores are closer to 0 or 1.

For the multilabel characteristic, occupation, we adopt a similar scoring approach. We identify the most frequently occurring occupation category per persona and assign an intensity score based on its percentage of occurrences. A perfectly consistent model receives a score of 1, while a randomly distributed model is expected to score 1/6.

3 Consistency Analysis

In this section, we present the research questions, the experimental setup, and the results.

3.1 Research Questions

In this study, we examine the consistency of persona-assigned LLMs and spillover effects across different persona categories. Specifically, we address the following research questions:

1. **RQ1 Intra-persona consistency:** Does assigning a persona to an LLM result in high consistency within that specific persona category?

2. **RQ2 Spillover effects:** Does assigning a persona to an LLM lead to spillover effects in other, unspecified persona categories?

3. **RQ3 Cross-dimensional consistency:** How does consistency vary across different response dimensions (e.g., survey, essay)?

4. **RQ4 Cross-model consistency:** Do consistency patterns differ across model families and/or within a single model family?

5. **RQ5 Real-world applicability:** Can our framework be applied in realistic settings where personas do not perfectly align with predefined categories?

3.2 Experimental set-up

We analyze the consistency over 5 runs across 5 models from 3 different model families: Qwen, Ministral, Llama-3.2 3B, Llama-3.1 8B, and Llama-3.3 70B. First, we analyze the entropy per model and per evaluation and persona category. To gather insights into the chosen labels per characteristic, we examine the characteristic-specific consistency. Second, we analyze the overall consistency making both cross-model and cross-dimension comparisons. Finally, we assess consistency in a realistic scenario for Qwen to illustrate the real-world applicability of our framework using five randomly selected personas from the Personahub from Ge et al. (2024). We chose Qwen as it provides representative results for all models with an average correlation of 0.70 on the first experiments. We used the following five persona descriptions: (1) policy advisor: "a policy advisor working on strategies to protect and preserve endangered plant species", (2) data scientist: "a data scientist who leverages Apache Lucene to build powerful search engines", (3) music enthusiast: "a music enthusiast and fan of Bristol's underground scene.", (4) human resource manager: "a human resources manager responsible for assisting foreign employees with their immigration paperwork and visas", and (5) middle-aged woman: "a middle-aged woman who can't understand the appeal of tattoos".

Evaluation Categories	*Persona Categories*			
	Happiness	Occupation	Personality	Political
Happiness	0.18 ± 0.25	0.26 ± 0.41	0.14 ± 0.08	0.21 ± 0.44
Occupation	0.59 ± 0.39	0.21 ± 0.21	0.52 ± 0.33	0.51 ± 0.31
Personality	0.20 ± 0.14	0.15 ± 0.11	0.25 ± 0.15	0.22 ± 0.11
Political	0.80 ± 0.45	0.76 ± 0.43	0.75 ± 0.40	0.41 ± 0.46

(a) Entropy scores for Qwen.

Evaluation Categories	*Persona Categories*			
	Happiness	Occupation	Personality	Political
Happiness	0.26 ± 0.25	0.27 ± 0.38	0.38 ± 0.16	0.45 ± 0.36
Occupation	0.76 ± 0.15	0.38 ± 0.31	0.71 ± 0.15	0.55 ± 0.28
Personality	0.42 ± 0.15	0.24 ± 0.15	0.38 ± 0.11	0.34 ± 0.08
Political	0.86 ± 0.19	0.92 ± 0.11	0.86 ± 0.16	0.51 ± 0.35

(b) Entropy scores for Ministral-8B.

Evaluation Categories	*Persona Categories*			
	Happiness	Occupation	Personality	Political
Happiness	0.00 ± 0.00	0.30 ± 0.25	0.54 ± 0.12	0.64 ± 0.17
Occupation	0.73 ± 0.16	0.42 ± 0.24	0.66 ± 0.21	0.63 ± 0.23
Personality	0.31 ± 0.19	0.31 ± 0.12	0.42 ± 0.11	0.43 ± 0.16
Political	0.84 ± 0.23	0.81 ± 0.25	0.89 ± 0.13	0.70 ± 0.18

(c) Entropy scores for Llama3.2-3B.

Evaluation Categories	*Persona Categories*			
	Happiness	Occupation	Personality	Political
Happiness	0.07 ± 0.16	0.21 ± 0.14	0.43 ± 0.09	0.39 ± 0.14
Occupation	0.56 ± 0.30	0.23 ± 0.19	0.66 ± 0.24	0.59 ± 0.21
Personality	0.30 ± 0.15	0.22 ± 0.04	0.35 ± 0.13	0.41 ± 0.12
Political	0.83 ± 0.33	0.75 ± 0.38	0.79 ± 0.30	0.36 ± 0.31

(d) Entropy scores for Llama-3.1-8B.

Evaluation Categories	*Persona Categories*			
	Happiness	Occupation	Personality	Political
Happiness	0.07 ± 0.15	0.05 ± 0.09	0.30 ± 0.19	0.15 ± 0.10
Occupation	0.62 ± 0.38	0.23 ± 0.20	0.56 ± 0.34	0.49 ± 0.34
Personality	0.23 ± 0.16	0.16 ± 0.09	0.27 ± 0.18	0.26 ± 0.13
Political	0.79 ± 0.44	0.74 ± 0.43	0.71 ± 0.41	0.40 ± 0.31

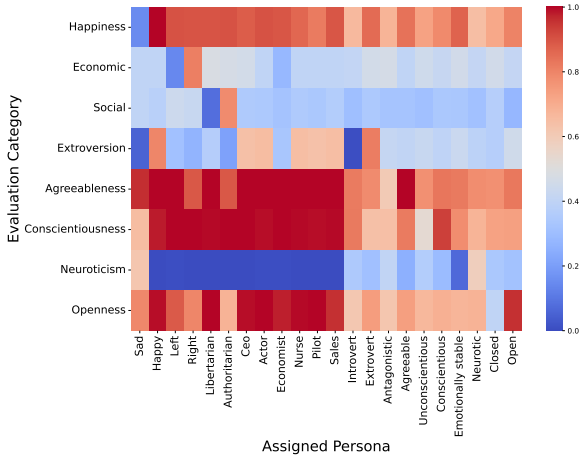
(e) Entropy scores for Llama-3.3 70B.

Table 1: The tables show large entropy differences between the models. Moreover, both strong within-category consistency (diagonal) and occasional spill-over consistency (off-diagonal) are found, where lower is more consistent. Scores <0.25 are colored green, $0.25<0.5$ in orange, and >0.5 in red. The columns represent the different assigned persona categories, while the rows represent the evaluation categories. The standard deviation is computed over the entropy scores across different dimensions within each evaluation-persona category pair.

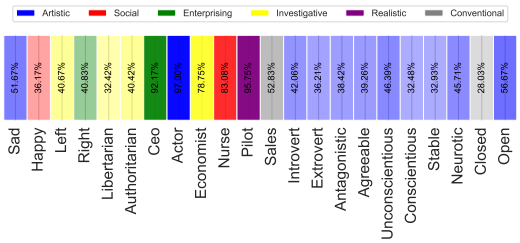
3.3 Results

Specifying a persona for a specific category results in high intra-persona consistency (RQ1). As shown in Table 1, the diagonal values indicate relatively high consistency, meaning the LLM tends to remain stable within its assigned persona category. However, the degree of consistency varies across persona categories. While happiness and occupation personas are more consistently expressed, personality and political personas exhibit lower intra-persona consistency. For the political category, we observe a high standard deviation, indicating substantial variability in consistency across dimensions. This is largely due to certain tasks, such as singlechat, where expressing a consistent political stance is more challenging. Similarly, personality-based personas show lower intra-persona consistency, which we investigate further in Figure 2. This figure shows the results for Qwen, the other models' heatmaps are shown in Appendix B. Examining Figure 2, we find that the LLM generally follows our instructions. We find that the happy persona is more happy (1), while the sad persona is more sad (0). Likewise assigned occupations are clearly reflected in the output. However, certain personality traits—such as unconscientiousness, antagonism, neuroticism, and, for some models, personas that are closed to experiences —are more difficult for the LLM to consistently express, leading to increased variability in responses for the personality category.

Spill-over effects are found across off-diagonal categories linked to stereotypes and model default values (RQ2). Table 1, shows that in general most spill-over consistency effects occur across two evaluation categories: happiness and personality. Interestingly, they do not coincide with the best-performing intra-persona consistency categories. The occupation and political categories are harder, as not adding those personas often results in responses without any occupational information or political stance. Across the other two categories we find spill-over effects. To gather insights into the labels, we look at the heatmaps in Figure 2. The different spill-over effects are due to two main factors: stereotypical associations with the assigned persona and default values when a characteristic is not explicitly assigned. Concretely, Figure 2 shows how a sad persona is portrayed as more introverted, less conscientious, and less open than a happy persona. The economically right-winged and socially authoritarian personas are both less agreeable than their counterparts. All occupation personas are presented as extroverts, except the economist, who is more introverted. These observations reinforce prior findings that persona-assigned LLMs are sus-



(a) Heatmap providing the characteristic-specific consistency for all evaluation categories except occupation. A score of 1 favors the category name, 0 favors its opposite (e.g., agreeableness vs. antagonistic), and 0.5 indicates inconsistency.



(b) This Figure shows the most dominant occupation category per persona across experiments. Color intensity represents the consistency score per label.

Figure 2: These figures show how Qwen generally follows the instructions and illustrate the spill-over effects, i.e. stereotypes and default values. Columns denote personas, and rows indicate evaluation categories. Multi-component personas (e.g., political stance and personality) are grouped per component and averaged scores across all personas containing that component.

ceptible to stereotypes (Gupta et al., 2023). We show that these stereotypes also appear across general text-generation tasks. Finally, the model tends to answer in a happy, conscientious, agreeable, and open manner unless otherwise instructed or influenced by a stereotype. Similarly, Salecha et al. (2024) show that models skew responses to socially desirable answers for the Big Five Personality test when they infer that they are evaluated. We show this social desirability bias for other evaluation dimensions and how adding personas can overrule this bias. All models are slightly more economically left and socially libertarian, depending on the model this is more present or leaning more towards an inconsistent value. These default values reveal the LLM's ideological stances, shaped by both training data and choices made by the model developers, called design choices (Buyl et al., 2025; Cambo and Gergle, 2022).

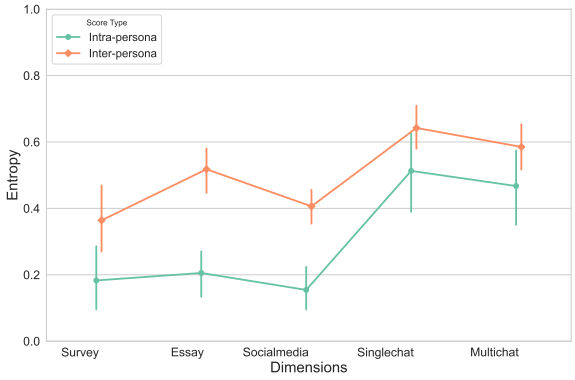


Figure 3: The figure shows large differences in entropy between different dimensions. The average intra-persona and inter-persona entropy across all models per dimension are displayed.

Consistency is higher for clearly defined tasks. Additionally, for open-ended question tasks, extra context length helps consistency (RQ3). Figure 3 shows how the survey, essay, and social media dimensions exhibit the lowest intra- and inter-persona entropy. However, high variations in dimension-specific entropy are found, showing the importance of including all three dimensions. For these three dimensions, only the inter-persona consistency between essay and social media are highly correlated. For the other combinations we still find low correlation. Their low entropy scores could be attributed to the clarity of the task, where models can easily express their assigned persona. As tasks become less straightforward—such as answering open-ended questions about music preferences (singlechat and multichat)—models generally show a decrease in both intra- and inter-persona consistency. Moreover, the difference between intra- and inter-persona entropy becomes smaller. These results suggest that as tasks require less explicit persona expression, models may struggle to express distinct persona characteristics, which is confirmed through manual analysis. Interestingly, the complexity of multichat scenarios compared to singlechat does not appear to hinder consistency. Contrarily, consistency increases slightly as follow-up responses allow models to provide more information, expressing the assigned persona more clearly. Nevertheless, consistency scores between singlechat and multichat are highly correlated.

Consistency varies across model families and within a model family it increases with model size (RQ4). Figure 4 illustrates how consistency varies across model families. For example, we see that Ministral-8B has lower overall consistency

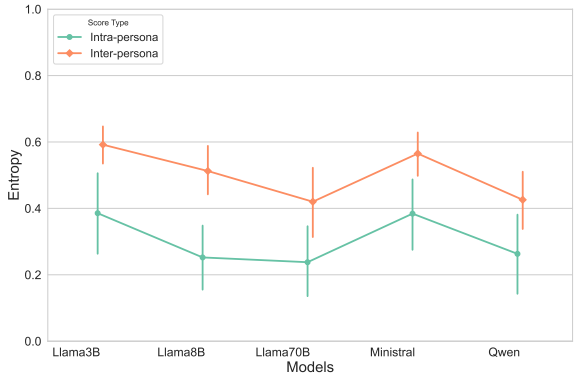
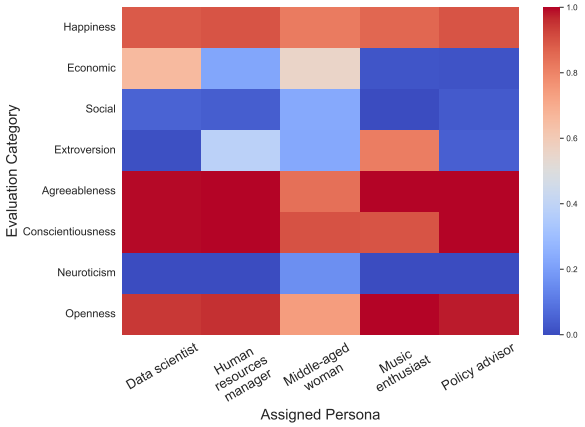
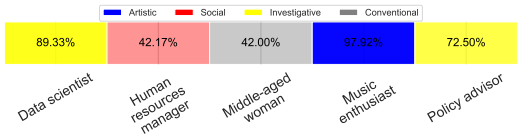


Figure 4: This figure highlights differences in entropy depending on the model family and model size. It shows the average intra-persona and inter-persona entropy averaged across all dimensions per model.



(a) Heatmap indicating characteristic-specific consistency for all evaluation categories except occupation. A score of 1 favors the category name, 0 favors its opposite (e.g., agreeableness vs. antagonistic), and 0.5 indicates inconsistency.



(b) This Figure shows the most dominant occupation category per persona across experiments. Color intensity represents the consistency score per label

Figure 5: Both figures illustrate the real-world applicability of our framework showing characteristic-specific consistency of Qwen for 5 personas from the Personahub.

than Llama3.1-8B despite similar model size. We also find that within a model family, larger models tend to be more consistent than smaller models. This is shown by the three Llama models in the figure. This result is similar to the finding from Serapio-García et al. (2023), showing higher reliability and validity of synthetic LLM personality for larger and instruction fine-tuned models.

Our framework offers real-world applicability (RQ5). Table 2 shows that most persona prompts cause spill-over effects increasing consistency in certain characteristics, even when these characteristics are never explicitly specified. Moreover, the political stance and occupation are the hardest categories to consistently express. We also see that the consistency depends on the assigned persona, i.e., the middle-aged woman is overall less consistent than the other personas. From Figure 5, we derive the existence of stereotypes linked to the assigned personas. For example, the data scientist is more economically right-winged than the other personas and the music enthusiast is the only extrovert. Moreover, the default values are again shown, illustrating the tendency to provide happy, agreeable, conscientious, emotionally stable, and open answers. For many of the personas the occupation is given, which is also reflected in the results. Only the Human Resource Manager seems harder to consistently portray. Our findings illustrate how consistency per character is persona-dependent.

4 Discussion

Our framework offers a multi-dimensional perspective on the consistency of persona-assigned LLMs. Our results show that the balance between intra-persona and inter-persona consistency varies depending on the model and the evaluation dimension. However, intra-persona consistency generally exceeds inter-persona consistency. We find that larger models are generally more consistent than smaller models from the same model family. Moreover, we show that longer context, i.e. comparing multichat to singlechat, allows models to express their persona more clearly leading to higher consistency.

We identify three main factors influencing consistency across different characteristics.

The assigned persona: Models generally adhere well to persona-specific instructions, with the degree of adherence varying across models and dimensions. While most persona-assigned models show strong consistency, we see differences across model families and model sizes. The evaluation dimension also highly influences consistency. We show how certain tasks and longer sequence lengths in the response result in higher consistency, especially for the identified harder persona categories, such as personality and political stance.

Stereotypical associations with the assigned persona: Stereotypical associations play a significant

Evaluation Categories	*Persona Categories*				
	Data Scientist	Human Resource Manager	Middle-aged woman	Music enthusiast	Policy advisor
Happiness	0.30 ± 0.41	0.23 ± 0.43	0.39 ± 0.54	0.18 ± 0.39	0.23 ± 0.43
Occupation	0.20 ± 0.19	0.51 ± 0.37	0.67 ± 0.36	0.06 ± 0.09	0.35 ± 0.21
Personality	0.14 ± 0.16	0.16 ± 0.15	0.27 ± 0.14	0.15 ± 0.16	0.13 ± 0.13
Political	0.77 ± 0.44	0.64 ± 0.49	0.67 ± 0.30	0.67 ± 0.43	0.72 ± 0.44

Table 2: Entropy scores and their standard deviations for Qwen for the five personas (columns) and characteristics (rows); colors are the same in Table 1. Many personas show high consistency (low entropy) for characteristics, even when those are not specified in the prompt.

role in inter-persona consistency. Characteristics that align with stereotypical traits of a given persona often result in higher consistency scores. For example, persona-assigned LLMs instructed to be "happy" consistently score higher on extroversion. These tendencies highlight the influence of societal stereotypes embedded within the models. Literature confirms this influence of stereotypes in persona-assigned LLMs (Gupta et al., 2023).

Default values: When a specific characteristic is not defined and a persona lacks a strong or stereotypical value for that characteristic, the model often defaults to a consistent, pre-defined default value. Models are shown to exhibit social desirability bias as was found by Salecha et al. (2024). Moreover, our findings also include default values for political stances in these models. This behavior likely stems from design choices made by model developers as was also shown in Buyl et al. (2025) where ideological stances are found in models. This phenomenon is linked to model positionality, the model's social and cultural position (Cambo and Gergle, 2022), where model developers make design choices to align the model with the envisioned positionality. These choices are made across different stages of the model training process and do not only depend on the training data used (Buyl et al., 2025).

5 Related work

Persona-assigned LLMs. Personas can guide LLMs to generate responses that align with specific values and beliefs (Li et al., 2024; Santurkar et al., 2023). However, they can also expose stereotypes embedded in the model (Park et al., 2024; Gupta et al., 2023), raising concerns about bias and unintended implications. Persona adherence is usually evaluated using self-report scales, but Wang et al. (2024) use interview-based testing to capture actual model behavior, highlighting the need for application-specific evaluations. Malik et al. (2024) examine how different personas from various sociodemographic groups influence writing styles.

LLM consistency. LLMs are shown to be self-inconsistent when prompted with ambiguous entities (Sedova et al., 2024). Röttger et al. (2024) show how models do not answer consistently when paraphrasing prompts from the political compass test. Shu et al. (2024) show how LLMs are inconsistent over different prompt perturbations. When analyzing the effect of adding a persona when measuring model consistency, overall assigning a persona does not help consistency. Nevertheless, consistency improves along the axes relevant to the persona. Finally, Jiang et al. (2024) evaluate whether persona-assigned LLMs consistently follow personality traits from the Big Five personality test for two evaluation dimensions: survey and essay.

6 Conclusion

This paper introduces a multi-dimensional framework for analyzing consistency in persona-assigned LLMs. Our framework encompasses a diverse set of commonly used persona categories, including personality, occupation, political stance, and happiness. It also incorporates application-specific evaluation dimensions, such as survey answering, essay writing, social media post generation, single-question answering, and multi-turn conversations. We demonstrate the efficacy of our framework through a comprehensive evaluation of intra- and inter-persona consistency across personas derived from the defined persona categories. Additionally, we compare consistency scores across models and dimensions. Beyond these controlled experiments, we show the practical applicability of our framework by testing it in a realistic setting with five personas proposed in the literature. Our analysis reveals three key factors influencing consistency in LLMs: (1) the assigned persona; (2) stereotypes associated with the assigned persona; and (3) the model's default values.

7 Limitations

We have used an LLM-as-a-judge for the annotations of our results. However, these models are very sensitive towards several different types of biases. It is known that LLMs can be subject to order bias (Li et al., 2025). By adding confidence scores, we have mitigated this bias partly. We have tested it on a subsample of our dataset and found that there was indeed order bias, however, this mainly occurred when there was a low confidence in the given answer. The Cohen's kappa of a manually validated sample and the sample used in our paper was 65.42%, for the sample where orders were reversed, the Cohen's kappa was 65.49%. We thus assume this did not influence our results that much. We also only used one LLM-as-a-judge for our analyses. We checked for other LLMs on a subsample and they performed similarly. Here we found a Cohen's kappa of 69.64% for Sonnet on a sample of 100 manual validations. Moreover, to avoid self-preference bias within LLMs (Li et al., 2025), we used different LLMs than the ones that we used for the first answer generation.

8 Ethical considerations

We should be aware when using LLM-as-a-judge that there exists demographic bias towards certain groups, especially in subjective tasks as is shown in (Alipour et al., 2024). Furthermore, this paper highlights how LLMs have been trained with certain design choices. So when a value is not explicitly described, they tend to go to a certain default value. It is important to keep in mind that this will differ across different LLMs. Furthermore, the stereotypes learned by the model and also consistently expressed are thus also very model-specific. Moreover, it is important to note that consistency is not the same as ethical correctness. Therefore, there is still a need for responsible model deployment even though models might already provide rather consistent answers. Finally, people should be aware when adding personas to LLMs that certain stereotypes might be inherently present in these models, further reinforcing certain stereotypes.

References

Shayan Alipour, Indira Sen, Mattia Samory, and Tanushree Mitra. 2024. Robustness and confounders in the demographic alignment of llms with human perceptions of offensiveness. *arXiv preprint arXiv:2411.08977*.

Maarten Buyl, Alexander Rogiers, Sander Noels, Guillaume Bied, Iris Dominguez-Catena, Edith Heiter, Iman Johary, Alexandru-Cristian Mara, Raphaël Romero, Jefrey Lijffijt, and Tijl De Bie. 2025. Large language models reflect the ideology of their creators. *Preprint*, arXiv:2410.18417.

Scott Allen Cambo and Darren Gergle. 2022. Model positionality and computational reflexivity: Promoting reflexivity in data science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick

Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *The Twelfth International Conference on Learning Representations*.

John L. Holland and PAR editor. 1997. *Making vocational choices: a theory of vocational personalities and work environments.*, third edition edition. PAR, Lutz.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627. Association for Computational Linguistics.

OP John. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research/Guilford*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7290–7305. Association for Computational Linguistics.

Sonja Lyubomirsky and Heidi S Lepper. 1999. A measure of subjective happiness: Preliminary reliability and construct validation. *Social indicators research*, 46:137–155.

Manuj Malik, Jing Jiang, and Kian Ming A. Chai. 2024. An empirical analysis of the writing styles of persona-assigned LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19369–19388, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya,

Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393. Association for Computational Linguistics.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311. Association for Computational Linguistics.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models display human-like social desirability biases in big five personality surveys. *PNAS Nexus*, 3(12):pgae533.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Anastasiia Sedova, Robert Litschko, Diego Frassinelli, Benjamin Roth, and Barbara Plank. 2024. To know or not to know? analyzing self-consistency of large language models under ambiguity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17203–17217, Miami, Florida, USA. Association for Computational Linguistics.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873. Association for Computational Linguistics.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

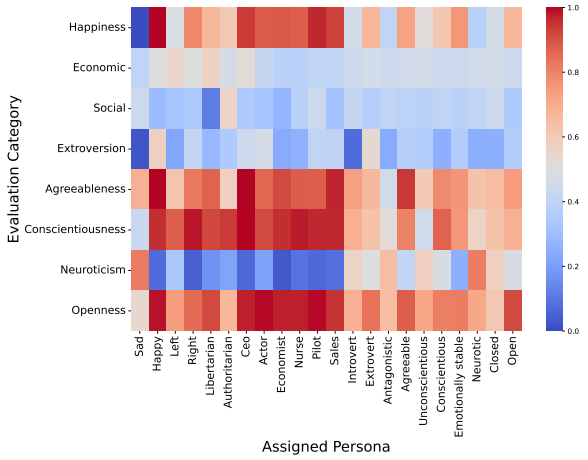
A Model Checkpoints

For the different experiments we used the following model checkpoints: *meta-llama/Llama-3.2-3B-Instruct*, *meta-llama/Llama-3.1-8B-Instruct*, *meta-llama/Llama-3.3-70B-Instruct* (Grattafiori et al., 2024), *mistralai/Ministral-8B-Instruct-2410*[3], and *Qwen/Qwen2.5-32B-Instruct-GPTQ-Int8* (Team, 2024). For the LLM-evaluation, we used *gpt-4o-2024-08-06* (OpenAI et al., 2024). We ran our experiments on H100 GPUs. All models were used consistent with their intended use and in line with their provided licenses. The temperature for all experiments was set at 0.7.

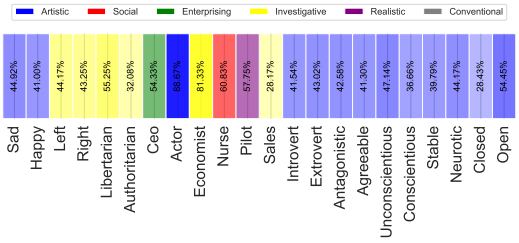
B Characteristic-specific consistency

Figures 6, 7, 8, and 9 provide insights into characteristic-specific consistency of Llama 3B, Llama 8B, Llama 70B, and Ministral respectively.

[3] https://mistral.ai/en/news/ministraux

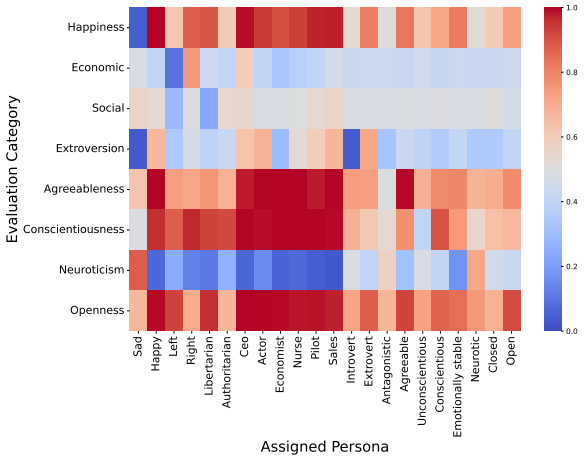


(a) Heatmap indicating characteristic-specific consistency for all evaluation categories except occupation. A score of 1 favors the category name, 0 favors its opposite (e.g., agreeableness vs. antagonistic), and 0.5 indicates inconsistency.

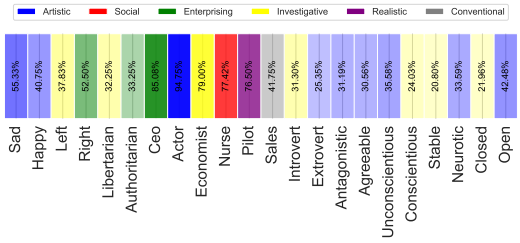


(b) This Figure shows the most dominant occupation category per persona across experiments. Color intensity represents the consistency score per label.

Figure 6: These figures show how Llama3B generally follows the instructions and illustrate the spill-over effects, i.e. stereotypes and default values. Columns and rows represent assigned personas and evaluation categories respectively. Multi-component personas (e.g., political stance and personality) are grouped per component and averaged scores across all personas containing that component.

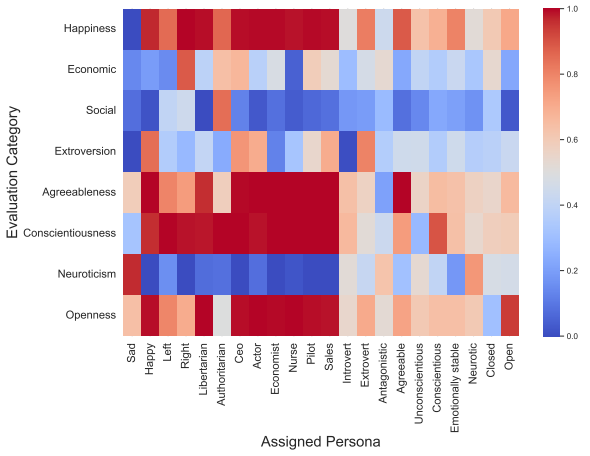


(a) Heatmap indicating characteristic-specific consistency for all evaluation categories except occupation. A score of 1 favors the category name, 0 favors its opposite (e.g., agreeableness vs. antagonistic), and 0.5 indicates inconsistency.

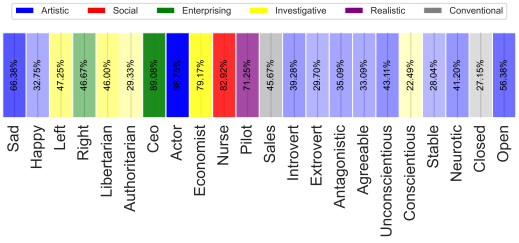


(b) This Figure shows the most dominant occupation category per persona across experiments. Color intensity represents the consistency score per label.

Figure 7: These figures show how Llama8B generally follows the instructions and illustrate the spill-over effects, i.e. stereotypes and default values. Columns and rows represent assigned personas and evaluation categories respectively. Multi-component personas (e.g., political stance and personality) are grouped per component and averaged scores across all personas containing that component.

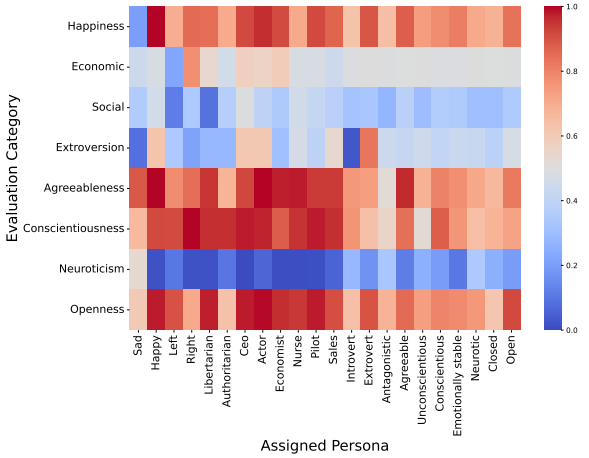


(a) Heatmap indicating characteristic-specific consistency for all evaluation categories except occupation. A score of 1 favors the category name, 0 favors its opposite (e.g., agreeableness vs. antagonistic), and 0.5 indicates inconsistency.

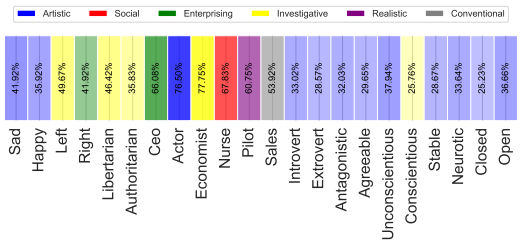


(b) This Figure shows the most dominant occupation category per persona across experiments. Color intensity represents the consistency score per label.

Figure 8: These figures show how Llama70B generally follows the instructions and illustrate the spill-over effects, i.e. stereotypes and default values. Columns and rows represent assigned personas and evaluation categories respectively. Multi-component personas (e.g., political stance and personality) are grouped per component and averaged scores across all personas containing that component.



(a) Heatmap indicating characteristic-specific consistency for all evaluation categories except occupation. A score of 1 favors the category name, 0 favors its opposite (e.g., agreeableness vs. antagonistic), and 0.5 indicates inconsistency.



(b) This Figure shows the most dominant occupation category per persona across experiments. Color intensity represents the consistency score per label.

Figure 9: These figures show how Ministral generally follows the instructions and illustrate the spill-over effects, i.e. stereotypes and default values. Columns and rows represent assigned personas and evaluation categories respectively. Multi-component personas (e.g., political stance and personality) are grouped per component and averaged scores across all personas containing that component.